Curtailment and Stochastic Curtailment to Shorten the CES-D

Finkelman, M.D.; Smits, N.; Kim, W.; Riley, B.

published in
Applied Psychological Measurement
2012

DOI (link to publisher)
10.1177/0146621612451647

document version
Publisher's PDF, also known as Version of record

Curtailment and Stochastic Curtailment to Shorten the CES-D

Matthew D. Finkelman, Niels Smits, Wonsuk Kim and Barth Riley

Applied Psychological Measurement published online 24 July 2012
DOI: 10.1177/0146621612451647

Article

Curtailment and Stochastic Curtailment to Shorten the CES-D

Applied Psychological Measurement
XX(X) 1–27

DOI: 10.1177/0146621612451647
http://apm.sagepub.com

Matthew D. Finkelman[1], Niels Smits[2], Wonsuk Kim[3], and Barth Riley[4]

Abstract

The Center for Epidemiologic Studies–Depression (CES-D) scale is a well-known self-report instrument that is used to measure depressive symptomatology. Respondents who take the full-length version of the CES-D are administered a total of 20 items. This article investigates the use of curtailment and stochastic curtailment (SC), two sequential analysis methods that have recently been proposed for health questionnaires, to reduce the respondent burden associated with taking the CES-D. A post hoc simulation based on 1,392 adolescents' responses to the CES-D was used to compare these methods with a previously proposed computerized adaptive testing (CAT) approach. Curtailment lowered average test lengths by as much as 22% while always matching the classification decision of the full-length CES-D. SC and CAT achieved further reductions in average test length, with SC's classifications exhibiting more concordance with the full-length CES-D than do CAT's. Advantages and disadvantages of each method are discussed.

Keywords

CES-D, curtailment, stochastic curtailment, computerized adaptive testing, respondent burden

Introduction

Rapid and accurate screening of depression has become a public health imperative. Comparison of two large U.S. representative adult surveys using structured clinical interview data revealed that depression rates increased from 3.33% in 1990-1991 to 7.06% in 2001-2002 (Compton, Conway, Stinson, & Grant, 2006). Lifetime prevalence of depression has been estimated at 15%

[1]Tufts University School of Dental Medicine, Boston, MA, USA
[2]Vrije Universiteit, Amsterdam, Netherlands
[3]Measured Progress, Dover, NH, USA
[4]University of Illinois College of Nursing, Chicago, IL, USA

Corresponding Author:
Matthew D. Finkelman, Tufts University School of Dental Medicine, 75 Kneeland Street, Room 105, Boston, MA 02111, USA
Email: matthew.finkelman@tufts.edu

(Kessler et al., 1994). A condition that is commonly seen by primary care physicians (Mitchell & Coyne, 2007; Spitzer et al., 1995; Wells, Sturm, Sherbourne, & Meredith, 1996), depression is the most costly disorder with respect to days lost to illness, impact on family and employer, and suicide risk, and is an independent predictor of mortality and cardiovascular disease (Glassman & Shapiro, 1998; Murray & Lopez, 1996; World Health Organization, 2001). Evidence suggests that depressed patients who are identified and receive appropriate treatment by primary care practitioners have the best outcomes (U.S. Preventive Services Task Force, 2002). Screening is recognized as an important first step in this process, and self-report instruments have generally been shown to be effective tools for screening (Mitchell & Coyne, 2007).

The Center for Epidemiologic Studies–Depression (CES-D; Radloff, 1977; M. M. Weissman, Sholomskas, Pottenger, Prusoff, & Locke, 1977) scale is a 20-item self-report instrument designed to measure depression in the general population. The CES-D is one of the most widely used depression instruments in research and has also been extensively used for depression screening. The scale resulted from a series of studies conducted by the center to develop techniques for the ongoing measurement of psychiatric impairment (M. M. Weissman et al., 1977). In its original validation study (Radloff, 1977), the CES-D exhibited good internal consistency (α of .80 or above across demographic groups) and moderate test–retest reliability (correlations of .40 or above).

Substantial research has been conducted to determine whether the CES-D and other self-report symptom measures can be used as the basis for clinical diagnosis (Breslau, 1985; Fechner-Bates, Coyne, & Schwenk, 1994; Myers & Weissman, 1980; Prescott et al., 1998; Roberts, Lewinsohn, & Seeley, 1991; Roberts & Vernon, 1983). The results generally indicate that although the CES-D is not sufficient to make a definitive diagnosis of depression, it can be used successfully as part of an initial depression screening process. In one study, it was revealed that using this instrument as the initial screening tool followed by a face-to-face diagnostic interview resulted in high predictive power to detect major depressive disorder but not dysthymia (Yang, Soong, Kuo, Chang, & Chen, 2004). Another study showed that the CES-D correlates with other self-report depression measures and clinician ratings as well as differentiating samples of individuals without depression from samples diagnosed with depression and related disorders (M. M. Weissman et al., 1977).

Although the CES-D has been used in many diverse applications (Sephton et al., 2009), the length of the 20-item version may limit its feasibility in some settings. Administering a large number of items leads to greater *respondent burden*, which can reduce the quality of a test taker's responses (Herzog & Bachman, 1980) or his or her willingness to take the questionnaire at all (Adams & Gale, 1982). Minimizing such respondent burden is a critical component of a questionnaire's design (Scientific Advisory Committee of the Medical Outcomes Trust, 2002). Completing the 20-item CES-D is typically manageable for healthy respondents, but it can be substantially more difficult for individuals who are elderly, are physically ill, or have problems with reading comprehension (Carpenter et al., 1998). One study reported that almost 10% of the elderly respondents at one site refused to answer all the CES-D items and that brevity is critical to reducing the refusal rate (Kohout, Berkman, Evans, & Cornoni-Huntley, 1993). Using shorter assessments is also important for minimizing dropout when respondents are assessed repeatedly over time (Kohout et al., 1993; Smits, Zitman, Cuijpers, den Hollander-Gijsman, & Carlier, 2012). Finally, questionnaires are often administered as a set, and the time difference between the abbreviated and full-length versions could be the determining factor in whether a measure of depressive symptoms is included in a lengthy set of instruments (Kohout et al., 1993). Therefore, several studies have investigated whether the CES-D can be shortened to facilitate its use. Evidence from these studies suggests that shorter versions of the CES-D can achieve adequate internal

consistency, sensitivity, and specificity (Carpenter et al., 1998; Cole, Rabin, Smith, & Kaufman, 2004; Grzywacz, Hovey, Seligman, Arcury, & Quandt, 2006; Poulin, Hand, & Boudreau, 2005) and discriminate between depressed and nondepressed subgroups (Santor & Coyne, 1997).

Using a short form is not the only way to reduce the respondent burden of a questionnaire. Gains in efficiency can also be obtained by exploiting modern advances in computer-based assessments. Perhaps the most well known of these advances is computerized adaptive testing (CAT), which has received much attention as a mode of administration for health instruments (Choi, Reise, Pilkonis, Hays, & Cella, 2010; Fries, Cella, Rose, Krishnan, & Bruce, 2009; Reeve et al., 2007; Smits, Cuijpers, & van Straten, 2011). Whereas traditional ''static'' test designs call for all respondents to receive the same set of items, CAT customizes the assessment at the individual level (Lord, 1980). In particular, the CAT paradigm prescribes that each respondent's items be based on his or her previous answers, so that the questionnaire can provide maximal information about the respondent taking it. The test is typically terminated when a fixed number of items is reached or when the measurement precision reaches a specified value. CAT has been an integral part of the Patient-Reported Outcomes Measurement Information System (PROMIS; Choi et al., 2010; Fries et al., 2009; Reeve et al., 2007) and has also been examined as a mode of administration for the CES-D (Smits et al., 2011). In a post hoc simulation of 1,392 adolescents, CAT was found to improve the efficiency of the CES-D when measuring depression along a continuous spectrum (Smits et al., 2011).

Another option to reduce respondent burden is to use curtailment or stochastic curtailment (SC), two sequential analysis techniques that were recently applied to health questionnaires (M. D. Finkelman, He, Kim, & Lai, 2011). Like CAT, curtailment and SC require that testing be conducted by computer so that interim analyses can be performed between items. Unlike CAT, however, these methods are used only for assessments that are designed to classify respondents into categories. Their strategy is to cease testing before ''unnecessary'' items—items that cannot or are unlikely to change the respondent's classification—are administered. In a post hoc simulation using data from the Medicare Health Outcomes Survey, curtailment and SC reduced the average number of items administered while maintaining equal or comparable classification accuracy (M. D. Finkelman et al., 2011). However, neither method has been investigated as a means of shortening the CES-D. Furthermore, as is explained in the section Application of CAT, Curtailment, and SC to the CES-D, the particular implementation of SC considered in M. D. Finkelman et al. was conservative in terms of its stopping rule; a more aggressive variation of this method could potentially yield greater improvements in average test length.

The objectives of this study are (a) to adapt previously proposed formulations of curtailment and SC to the CES-D, (b) to introduce a more aggressive version of SC that can be used for the CES-D as well as for other health questionnaires, and (c) to compare the curtailment and SC approaches with CAT, based on post hoc simulation of the same data set that was used in Smits et al. (2011).

The remainder of the article is organized as follows. The section Traditional Scoring of the CES-D provides a brief overview of how the full-length CES-D is traditionally scored. The section Application of CAT, Curtailment, and SC to the CES-D reviews the CAT-based CES-D proposed in Smits et al. (2011) and adapts curtailment and SC to the CES-D. The more aggressive version of SC is also introduced. The sections Simulation Design and Results present the design and results of the post hoc simulation study. The section Summary and Discussion offers concluding remarks.

Table 1. The 20-Item CES-D

Item number	Item stem
1	I was bothered by things that usually don't bother me
2	I did not feel like eating; my appetite was poor
3	I felt that I could not shake off the blues even with help from my family or friends
4	I felt I was just as good as other people[a]
5	I had trouble keeping my mind on what I was doing
6	I felt depressed
7	I felt that everything I did was an effort
8	I felt hopeful about the future[a]
9	I thought my life had been a failure
10	I felt fearful
11	My sleep was restless
12	I was happy[a]
13	I talked less than usual
14	I felt lonely
15	People were unfriendly
16	I enjoyed life[a]
17	I had crying spells
18	I felt sad
19	I felt that people disliked me
20	I could not get going

Note: CES-D = Center for Epidemiologic Studies–Depression.
[a]Items 4, 8, 12, and 16 are scored in reverse of the other items.

Traditional Scoring of the CES-D

Table 1 shows the 20 items that make up the full-length CES-D. Each item asks the respondent to answer a question about his or her status during the past week. For the 16 items whose stems indicate greater depression, the original scoring rules (Radloff, 1977) are as prescribed below:

- Endorsement of ''rarely or none of the time (less than 1 day)'' results in a score of 0.
- Endorsement of ''some or a little of the time (1-2 days)'' results in a score of 1.
- Endorsement of ''occasionally or a moderate amount of time (3-4 days)'' results in a score of 2.
- Endorsement of ''most or all of the time (5-7 days)'' results in a score of 3.

For the four items whose stems indicate less depression (Items 4, 8, 12, and 16), ''reverse scoring'' is performed. That is, an answer of ''most or all of the time'' is scored 0, an answer of ''occasionally or a moderate amount of time'' is scored 1, an answer of ''some or a little of the time'' is scored 2, and an answer of ''rarely or none of the time'' is scored 3. The respondent's total number-correct score is obtained by summing the 20 item scores; a total score of $\geq$16 is typically used as the cutoff for clinical depression, although this value has been acknowledged by Radloff (1977) as arbitrarily selected. Some researchers have asserted that a $\geq$16 cutoff is too liberal, particularly when applied to adolescents and primary medical care patients (Santor & Coyne, 1997). Optimal cutoff values of $\geq$29, $\geq$24, and $\geq$22 have been found in different adolescent samples (Chabrol, Montovany, Chouicha, & Duconge, 2002; Cuijpers, Boluijt, & van Straten, 2008; Mojarrad & Lennings, 2002). The study that proposed a $\geq$22 cutoff (Cuijpers et al., 2008) used the same sample of adolescents that was used to develop

the CAT-based CES-D (Smits et al., 2011); this sample is also examined in the sections Simulation Design and Results of the present study.

Application of CAT, Curtailment, and SC to the CES-D

CAT

Item response theory (IRT). Before one implements CAT in an operational setting, it is typically necessary to fit a probabilistic model relating respondents' answers to the trait being measured by the questionnaire. This is most often done using the framework of IRT. In IRT, the trait being measured is considered to be a latent variable, usually referred to as θ. Respondents' answers are manifest variables; a given respondent's answer to item j will be denoted u_j (the different respondents could be indexed by i, but such notation is often suppressed for simplicity).

There are a number of IRT models that are available to quantify the relationship between θ and u_j (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985; Lord, 1980). A first step in choosing the appropriate model for a given questionnaire is to determine whether its items are *dichotomous* (yes or no), *polytomous* (more than two possible answers), or a mix of the two types. As described in the section Traditional Scoring of the CES-D, each of the CES-D's 20 items has four possible answers; hence, all of its items are polytomous. Smits et al. (2011) considered two polytomous IRT models for application to the CES-D: the graded response model (GRM; Samejima, 1969) and the partial credit model (PCM; Muraki, 1992). The authors ultimately chose to use the GRM due to its interpretability and ease of understanding. Application of the GRM to the CES-D is explained in detail in Smits et al.; briefly, this model begins by specifying a set of thresholds separating ''lower'' answers from ''higher'' answers. Every CES-D item has three thresholds: one separating a score of 0 from the scores 1 to 3, one separating 0 to 1 scores from 2 to 3 scores, and one separating 0 to 2 scores from a score of 3. The probability that a respondent with latent trait θ would score above a given threshold is modeled by a logistic curve. In particular, the probability that such a respondent would score above threshold v on item j is given as

$$P^*_{jv}(\theta) = \left[1 + \exp\left(-a_j\theta + b_{jv}\right)\right]^{-1}. \quad (1)$$

Here, a_j is the so-called discrimination parameter of item j. It relates to the item's ability to discern between respondents who have different θ values. b_{jv} is a ''difficulty'' parameter that is specific to the vth threshold of the item. The probability that a respondent with latent trait θ would receive a particular score can be obtained by subtracting adjacent values of $P^*_{jv}(\theta)$.

The parameters of each item are commonly estimated from respondent data using a statistical method such as joint maximum likelihood, conditional maximum likelihood, or marginal maximum likelihood (Embretson & Reise, 2000). With the resulting values treated as fixed, known quantities, the *person parameters* of new respondents can be estimated as well. In particular, each respondent's set of answers produces an estimate $\hat{\theta}$ of his or her underlying θ, typically using maximum likelihood, expected a posteriori (EAP), or maximum a posteriori (MAP) estimation (Embretson & Reise, 2000). The current paper focuses on the latter because it was used by Smits et al. (2011) in their development of a CAT-based CES-D.

To use MAP estimation, a Bayesian prior distribution $\pi_0(\theta)$ must first be specified along the θ scale. The standard normal distribution is a common choice for $\pi_0(\theta)$ and was utilized in previous CES-D research (Smits et al., 2011). After the respondent completes the questionnaire (which is assumed to be made up of N items), the likelihood function $L(\theta; \{u_j\}_{j=1}^{N})$ is computed. Under the usual IRT assumption of local independence (conditional independence of answers, given θ; Hambleton, Swaminathan, & Rogers, 1991), the likelihood function is equal to

$$L\left(\theta; \{u_j\}_{j=1}^{N}\right) = \prod_{j=1}^{N} P\left(U_j = u_j|\theta\right). \quad (2)$$

Here, U_j is a random variable representing the respondent's answer to item j; as described previously, u_j is the realized value of this random variable. $P(U_j = u_j|\theta)$ thus represents the probability that a respondent with a latent trait of θ would give the answer to item j that was actually observed, according to the GRM. The probabilities are multiplied due to the local independence assumption, yielding the likelihood function for the entire set of responses. θ values with higher likelihoods are thought to be more consistent with the respondent's answers than are θ values with lower likelihoods.

Once the likelihood function has been calculated, the posterior density of θ can be obtained. Again assuming that all N items have been administered, the density of a given value θ' is

$$\pi_N(\theta') = \frac{\pi_0(\theta')L\left(\theta'; \{u_j\}_{j=1}^{N}\right)}{\int \pi_0(\theta)L\left(\theta; \{u_j\}_{j=1}^{N}\right)d\theta}. \quad (3)$$

The value maximizing this function is the MAP estimate of θ and may be denoted $\hat{\theta}_N^{\mathrm{MAP}}$.

Like most IRT models, the GRM makes certain assumptions that must be satisfied for its use to be warranted. First, the assumption of *local independence* has already been mentioned. Second, the related assumption of *unidimensionality* states that there is a single or dominant factor that governs respondents' answers to all items (Hambleton et al., 1991). Third, it is assumed that items behave in a *monotone* pattern, that is, that as the level of the trait increases, the probability of exceeding each threshold also increases (Mokken, 1971). Before adopting the GRM as the model of choice, practitioners should also check that items do not exhibit differential item functioning (DIF; H.-H. Chang, Mazzeo, & Roussos, 1996; Crane, Gibbons, Jolley, & van Belle, 2006; Embretson & Reise, 2000; Holland & Wainer, 1993) and that the overall model fit is adequate (Embretson & Reise, 2000; McKinley & Mills, 1985). Each of these issues was considered in Smits et al. (2011), who concluded that the GRM is suitable for the CES-D.

From IRT to CAT. A well-known benefit of IRT is that respondents' scores can be placed on the same scale even when those respondents have not been administered the same set of items (Hambleton et al., 1991). This property of IRT facilitates the use of CAT, which involves the selection of different items for different respondents. It also allows *interim* $\hat{\theta}$ estimates to be obtained for each respondent when test administration is performed via computer. Interim estimates refer to ML, EAP, MAP, or other estimates that are calculated as the respondent is taking the questionnaire, using the partial answer string that has been observed up to the current time point. As explained in the following, such estimates are used by the CAT to determine which item, if any, should be presented next.

Suppose that the assessment of a given respondent is in progress, with $k < N$ items having been administered thus far. Let $\hat{\theta}_k$ denote the interim estimate based on the respondent's answers to these items. The first decision to be made is whether to continue testing (i.e., present another item) or halt the questionnaire (i.e., cease assessment and report the final results). One common approach is to stop examination if and only if a desired level of measurement precision has been reached (Smits et al., 2011; Thissen & Mislevy, 2000). Measurement precision is typically quantified via the standard error (*SE*), which is a function of θ. After each item has been administered, the *SE* function is evaluated at the value $\hat{\theta}_k$. If the resulting number is less than a prespecified ε, the stopping rule is invoked. Otherwise, the assessment continues, and a second

decision must be made: which item to administer at the next stage of the test. A popular method is to select the item that exhibits the greatest *Fisher information*, among all items that have not yet been presented to the respondent (Smits et al., 2011; Thissen & Mislevy, 2000). Like the *SE*, the Fisher information is a function of θ and is typically evaluated at $\hat{\theta}_k$. Intuitively, the higher an item's Fisher information is at $\hat{\theta}_k$, the greater its ability to discern between values near the respondent's interim estimate. See Dodd, De Ayala, and Koch (1995) and Samejima (1969) for formulas relating to Fisher information and the GRM.

CAT for classification. As mentioned in the Introduction, Smits et al. (2011) applied CAT to the CES-D, focusing primarily on the *estimation* of θ along a continuous spectrum. In addition to estimating θ, practitioners may wish to use the CES-D to *classify* respondents as either ''at risk'' or ''not at risk'' for depression. An obvious approach is to flag a respondent as ''at risk'' if and only if his or her estimated θ value meets or exceeds a prespecified cutoff along the latent trait continuum. More formally, let K denote the total number of items administered to a given respondent; if stopping is based on the *SE* criterion, K will vary by respondent (as in the section ''Item Response Theory,'' the indexing of different respondents by i is suppressed for simplicity). Let $\hat{\theta}_K$ denote the respondent's estimated θ value after testing has completed, and let θ^* denote the cutoff value. Then the respondent is flagged as ''at risk'' if and only if $\hat{\theta}_K \geq \theta^*$.

It is noteworthy that the preceding classification rule only considers whether the respondent's θ estimate meets or exceeds the cutoff value; it does not explicitly take into account the uncertainty associated with that estimate. By definition, all respondents who experience early stopping have θ estimates with *SE*s less than ε; however, some respondents who receive all 20 items may have *SE*s that are nontrivially higher than that value. One possible course of action would be to report ''no decision'' for respondents who have substantial uncertainty associated with their classification statuses. To be consistent with previous literature on computerized classification (e.g., Eggen & Straetmans, 2000; Lewis & Sheehan, 1990; Spray & Reckase, 1996; Vos, 2000); however, the preceding procedure makes a decision for every respondent. This approach is appropriate in the present context, wherein the CES-D is used as a screener and all respondents are either to be flagged or not flagged. Such classification of every respondent into one of two categories is consistent with standard practice in CES-D research (e.g., Pandya, Metz, & Patten, 2005; Radloff, 1977).

The classification method described in this section falls under the umbrella of computerized classification testing (CCT), which has been a fertile area of research over the last several decades. One of the most well-known procedures that can be used in CCT is the sequential probability ratio test (SPRT), which was originally developed by Wald (1947) in the sequential analysis context and was applied to assessment by Reckase (1983). Eggen (1999), Eggen and Straetmans (2000), and A. Weissman (2007) investigated different item selection criteria for the SPRT as well as extensions to cases with more than two possible classification decisions. Several of the item selection criteria were based on the concept of Kullback–Leibler information, which was originally proposed for computerized tests by H.-H. Chang and Ying (1996). Aside from the SPRT, methods for CCT include Bayesian decision theory (Lewis & Sheehan, 1990; Rudner, 2009; Vos, 2000), IRT-based confidence intervals (Thompson, 2007; Weiss & Kingsbury, 1984), and the generalized likelihood ratio test (Thompson, 2011). These procedures have different rules for when to stop examination and make a classification decision. In this study, attention focuses on the $SE < \varepsilon$ stopping criterion because it has been applied specifically to the CES-D in previous work (Smits et al., 2011).

One manner by which the different CCT methods can be categorized is by whether they specify an *indifference region* along the θ scale. Wald (1947) described the indifference region as the set of parameter values for which neither classification decision is strongly preferred over the other. Under an IRT framework, this region is typically an interval of θ values centered

around the cutoff θ^*. Points within the region may be viewed as the θ values that are so close to the cutoff that neither classification would constitute a grave error (Bartroff, Finkelman, & Lai, 2008; Y.-C. I. Chang, 2005; Thompson, 2011). Within CCT, indifference regions have primarily been utilized alongside the SPRT and some formulations of the generalized likelihood ratio test (Thompson, 2011). The simple classification rule defined above (namely, flagging respondents if and only if $\hat{\theta}_K \geq \theta^*$) allows the classification of each respondent without the specification of an indifference region.

Note that for some applications, a standard cutoff may be readily available along the number-correct score scale but not the θ scale (Hambleton et al., 1991). In these cases, the number-correct score cutoff can be converted to a corresponding θ^* value through the so-called test characteristic curve (TCC; Hambleton & de Gruijter, 1983; Hambleton et al., 1991). The TCC gives the expected number-correct score for each θ by summing the conditional expectations of all N individual items:

$$E(X_N|\theta) = \sum_{j=1}^{N} E\left(U_j|\theta\right). \tag{4}$$

Here, $E(U_j|\theta)$ represents a respondent's expected score on item j, given a latent trait of θ; this value is computed directly from the item's IRT parameters. X_N denotes the respondent's total number-correct score when all N items are administered, and $E(X_N|\theta)$ denotes its conditional expectation given θ. To calculate the θ^* value corresponding to a cutoff of X^* on the number-correct score scale, the inverse transformation of Equation 4 is used:

$$\theta^* = \{\theta : E(X_N|\theta) = X^*\}. \tag{5}$$

Because mild regularity conditions ensure that the TCC is a strictly increasing function of θ, Equation 5 provides a unique cutoff along the θ scale. The $\hat{\theta}_K \geq \theta^*$ criterion described above can then be used to classify each respondent.

Curtailment

As mentioned in the Introduction, CAT is an established tool that has been studied in numerous health-related applications, including the PROMIS initiative (Choi et al., 2010; Fries et al., 2009; Reeve et al., 2007) and the CES-D (Smits et al., 2011). However, it is by no means the only approach to enhancing the efficiency of measurement through computer-based testing. Techniques from the sequential analysis literature may also be examined in the pursuit of computerized questionnaires that exhibit low respondent burden and competitive diagnostic accuracy. One such technique that is simple to understand and apply is the method of curtailment (M. D. Finkelman et al., 2011).

As in the section ''CAT for Classification,'' suppose that each respondent is to be classified into one of the multiple (often two) mutually exclusive categories. A curtailment rule then prescribes that a respondent's questionnaire be stopped as soon as his or her classification decision can be known based on his or her previous answers. In other words, a curtailment rule would halt administration if every set of future responses would yield the same classification, given the answers up to the current point. The test resulting from this rule is called the *curtailed version* of the original test. The curtailed version always makes the same classification as the original test, but it may be shorter. Note that curtailment does not require the complicated framework of IRT and CAT; it is applicable even when simple classification rules are used and the ordering of items is identical for all respondents.

As an illustrative example, assume that a CES-D respondent is taking the instrument by computer, with items administered in the order that they are listed in Table 1. Furthermore, assume that scoring is performed as in the section ''Traditional Scoring of the CES-D,'' with a score of at least 16 required for the respondent to be flagged as ''at risk'' for depression. If the respondent's scores to the first 10 items are 2, 3, 0, 1, 2, 1, 2, 0, 2, and 3, then his or her cumulative score for those items is equal to 16. Because negative item scores do not exist for the CES-D, the respondent's final score for all 20 items is guaranteed to be at least 16; all possible future answers would result in an ''at risk'' classification. Noting this, a curtailment rule would cease assessment after 10 items and report that the respondent is flagged. Conversely, suppose that the respondent's cumulative score is 9 after the administration of 18 items. Because the maximum item score for the CES-D is 3, the respondent's final score cannot exceed 15, and thus a ''not at risk'' classification is inevitable. Rather than presenting the final two items, a curtailment rule would cease the assessment after 18 items and report that the respondent is not flagged.

If the traditional scoring method (section Traditional Scoring of the CES-D) is applied to the CES-D, curtailment's stopping rules and classification decisions can be stated formally as follows. Let X_k denote a respondent's cumulative score after k items. Let X^* denote the smallest total number-correct score, for all 20 items, that would result in flagging the respondent as ''at risk'' (X^* was specified as 16 in the preceding paragraph). At stage $k < 20$, a curtailment rule stops testing and flags the respondent if $X_k \geq X^*$, it stops testing and does not flag the respondent if $X_k + 3(20 - k) < X^*$, and it continues testing if neither of these conditions is satisfied. The $3(20 - k)$ term is included in the preceding formula because it is the maximum number of points that can be obtained during the remainder of the test (i.e., from item $k + 1$ to Item 20); if this value plus the current score X_k does not reach X^*, a ''not at risk'' classification is inevitable. At stage $k = 20$, testing always stops; the respondent is flagged if and only if $X_{20} \geq X^*$.

Theoretical properties of curtailment have been studied in the statistical literature (Eisenberg & Ghosh, 1980; Eisenberg & Simons, 1978). This sequential method has also been examined in the context of health questionnaires, using the Medicare Health Outcomes Survey as an example (M. D. Finkelman et al., 2011). However, no previous study has investigated it as a means of improving the efficiency of the CES-D.

SC

Motivation. The basic idea of SC is the same as that of curtailment: In both cases, future observations are examined with respect to their impact on the final classification decision. In SC, however, early stopping occurs not only when a respondent's classification is *certain* from a respondent's previous answers, but it also occurs when the classification is adequately *probable*. Consider an assessment, such as the CES-D, whose goal is to classify respondents into one of two mutually exclusive categories (e.g., ''at risk'' or ''not at risk'') and again suppose that this assessment is being administered by computer. Let γ_0 and γ_1 be two constants that are greater than 0.5. At each stage of testing, the probability that the respondent will ultimately be flagged as ''at risk'' by the full-length instrument is estimated. If this probability is greater than or equal to γ_1, the questionnaire is stopped and an ''at risk'' classification is reported. If the probability is less than or equal to $1 - \gamma_0$, the questionnaire is stopped and a ''not at risk'' classification is reported. If neither of these conditions is in effect, assessment continues. The test resulting from this rule is called the *stochastically curtailed version* of the original test.

SC is most famous for its use in the early stopping of clinical trials (Davis & Hardy, 1994; Lan, Simon, & Halperin, 1982; Leung, Wang, & Amar, 2003; Snapinn, Chen, Jiang, & Koutsoukos, 2006). It has recently been extended to the realm of individual assessment, first in

educational testing (M. Finkelman, 2008; M. D. Finkelman, 2010) and then in health questionnaires (M. D. Finkelman et al., 2011). However, like curtailment, it has never before been studied as a procedure to improve the efficiency of the CES-D.

The most difficult part in applying SC is to estimate the probability that an ''at risk'' classification will be made by the full-length instrument. A previously proposed method for the estimation process is adapted to the CES-D in the next section. As will be seen, this method is conservative in its approach to early stopping; a new method, designed to provide greater reduction in the number of items administered, is introduced in the section ''A Logistic Regression Approach to SC.'' Note that like the IRT-based classification procedure described in the section ''CAT for Classification,'' the curtailment and SC methods considered herein do not use an indifference region.

A previous formulation, adapted to the CES-D. Suppose that scoring of the CES-D is conducted as described in the section Traditional Scoring of the CES-D: Each item is scored on a scale from 0 to 3, with the total number-correct score defined as the sum of the individual item scores. As in the section ''Curtailment,'' the smallest total number-correct score that would result in flagging the respondent as ''at risk'' for depression is denoted by X^*. Assume the existence of a training data set that contains past respondents' answers to all 20 CES-D items. This data set is used to help determine whether a new respondent's examination will be terminated early.

Consider such a new respondent who is taking the CES-D via computer and has been administered $k < 20$ items so far. The SC procedure must decide whether to continue assessment or to cease testing in favor of an immediate classification. Its first step is to check whether the examination would be halted by the curtailment stopping rule outlined in the section ''Curtailment.'' If so, SC terminates the assessment and makes the same classification decision as would be made by the curtailment rule. If not, SC may still halt the assessment depending on the estimated probability that the respondent will be flagged as ''at risk'' by the full-length instrument. Applying a previous formulation of the method, called *stochastic curtailment via empirical proportions* for general health questionnaires (M. D. Finkelman et al., 2011), would result in the following steps for the CES-D:

1. Create two data sets, one containing the item answers of training-set respondents who were flagged as ''at risk'' by the full-length CES-D and the other containing the item answers of training-set respondents who were not flagged as ''at risk'' by the CES-D. These data sets will be referred to as T^+ and T^-, respectively. As always, respondents are flagged by the full-length CES-D if and only if their total number-correct score is at least X^*.
2. For each past respondent in T^+, determine the classification decision that would occur if the past respondent's answers to items $k+1$ through N were appended to the new respondent's answers to Items 1 through k. Let $\hat{P}_k^+$ denote the proportion of such decisions that are ''at risk'' classifications.
3. Analogously, for each past respondent in T^-, determine the classification decision that would occur if the past respondent's answers to items $k+1$ through N were appended to the new respondent's answers to Items 1 through k. Let $\hat{P}_k^-$ denote the proportion of such decisions that are ''at risk'' classifications.
4. If $\hat{P}_k^+ \geq \gamma_1$ and $\hat{P}_k^- \geq \gamma_1$, testing halts and the new respondent is flagged as ''at risk.''
5. If $\hat{P}_k^+ \leq 1 - \gamma_0$ and $\hat{P}_k^- \leq 1 - \gamma_0$, testing halts and the new respondent is not flagged as ''at risk.''
6. If testing is not halted in either Step 4 or 5, another item is presented.

As in curtailment, testing always stops if the 20th item is reached; in this case, the new respondent is flagged if and only if $X_{20} \geq X^*$.

The logic of the preceding procedure is to first assume that the new respondent's future answers will be similar to those of past respondents who received an ''at risk'' classification (Step 2). Next, it is assumed that the new respondent's future answers will be similar to those of past respondents who received a ''not at risk'' classification (Step 3). Testing is only halted if the probability of an ''at risk'' classification for the new respondent is at least γ_1, or at most $1 - \gamma_0$, under *both* of these assumptions. Therefore, when γ_0 and γ_1 are set to 0.95, as has been done previously (M. D. Finkelman et al., 2011), SC via empirical proportions is a conservative stopping rule. A new formulation of SC, presented in the following section, seeks to achieve lower average test lengths by avoiding such conservatism.

As in the previous section, assume the following.

- Scoring of the CES-D is conducted as described in the section Traditional Scoring of the CES-D.
- X^* denotes the smallest total number-correct score that would result in flagging the respondent as ''at risk'' for depression.
- A training data set containing past respondents' answers to all 20 CES-D items is on hand.

This section will develop a variation on SC that still stops whenever the curtailed version does, but otherwise differs from the stopping rules presented in the section ''A Previous Formulation, Adapted to the CES-D.''

Again suppose that a new respondent is taking the CES-D via computer and has been administered $k < 20$ items so far, resulting in a cumulative score of X_k. If the stopping rule of curtailment is invoked, testing halts and the classification decision is made as in the section ''Curtailment.'' Otherwise, a logistic regression is used to estimate the probability that the full-length instrument will make an ''at risk'' classification for this respondent. Specifically, a simple logistic regression model is fitted to the training data set; the independent variable is the cumulative score for Items 1 through k, and the dependent variable is the classification based on the full-length CES-D (0 = *not at risk*, 1 = *at risk*). The independent variable can be computed directly for all members of the training data set by summing their first k item scores; the dependent variable can be obtained by calculating each subject's total number-correct score and comparing with X^*. Members of the training set whose cumulative scores for Items 1 through k are greater than or equal to X^*, or strictly less than $X^* - 3(20 - k)$, are excluded when fitting the logistic regression. Such subjects' cumulative scores are high or low enough that their classification statuses are deterministic after item k; hence, their inclusion would be at odds with the notion of a probabilistic model.

Once the simple logistic regression model has been fitted to the appropriate training data, the probability that a new respondent will be classified as ''at risk'' by the full-length CES-D is estimated. Let $\hat{\alpha}_k$ denote the intercept of the logistic regression predicting classification status from X_k; let $\hat{\beta}_k$ denote the slope of this model. Then the estimated probability that the new respondent will be classified as ''at risk'' is given by

$$\hat{P}_k^l = \frac{\exp\left(\hat{\alpha}_k + \hat{\beta}_k X_k\right)}{1 + \exp\left(\hat{\alpha}_k + \hat{\beta}_k X_k\right)}. \quad (6)$$

If $\hat{P}_k^l \geq \gamma_1$, the test is stopped and an ''at risk'' classification is made; if $\hat{P}_k^l \leq 1 - \gamma_0$, the test is stopped and a ''not at risk'' classification is made; if neither of these conditions holds (and curtailment is not invoked), the test continues.

The preceding procedure involves a sequence of logistic regression models indexed by k. The fitting of a model after each stage of the test—for every new respondent—would be a computationally intensive task. Instead, all calculations are done ahead of time, that is, between the collection of the training data and the examination of the first new respondent. In particular, analysis of the training data is performed to determine (a) the smallest value of X_k sufficient for early stopping at time k in favor of an ''at risk'' classification and (b) the largest value of X_k sufficient for early stopping at time k in favor of a ''not at risk'' classification. Whether performing SC via empirical proportions or logistic regression, such ''stopping boundaries'' can be compiled in a simple lookup table and used for new respondents.

The SC method proposed in this section only requires that one estimated probability, $\hat{P}_k^l$, be unduly high or low for early stopping to occur. Such a requirement is generally less stringent than SC via empirical proportions, which requires $\hat{P}_k^+$ and $\hat{P}_k^-$ to be extreme for testing to cease. If γ_0 and γ_1 are held constant and both procedures are implemented, SC via logistic regression can be expected to produce smaller average test lengths, possibly accompanied by a decrement in classification accuracy. The magnitude of difference between the two procedures, as well as all other procedures described earlier, is examined in the following section using post hoc simulation of real data.

Simulation Design

The data set used in this study contained item responses from 1,392 Dutch adolescents who took the full-length version of the CES-D. Information about the sampling procedure has been reported in Cuijpers et al. (2008) and Smits et al. (2011); see these articles for full details. Briefly, the adolescents who participated in this study were recruited either in their secondary school or via the Internet. Ages ranged from 12 to 17, with a mean of 15.2 and a standard deviation (*SD*) of 1.0. The majority of the sample (63.6%) was female. In addition to taking the CES-D, a subset (N = 242) took part in a Mini-International Neuropsychiatric Interview (M.I.N.I.) to diagnose their level of depression (major, minor, or neither) via the criteria of the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; *DSM-IV*; American Psychiatric Association, 1994) and *International Classification of Diseases and Related Health Problems, 10th Revision* (ICD-10; Sheehan et al., 1998). Administered by telephone in this study, the M.I.N.I. is a structured interview that has been found to be reliable and valid for adolescents (Chabrol et al., 2002).

The purpose of the simulation was to compare the aforementioned adaptive and sequential methods to one another, as well as to the full-length CES-D, in terms of their classification properties and levels of respondent burden. Because responses to all 20 items had previously been collected for every respondent, the simulation was performed post hoc, with each method applied as if answers were being collected one at a time. The following approaches were compared:

- The full-length test. This procedure used all 20 CES-D items to classify each respondent. Scoring was performed as in the section ''Traditional Scoring of the CES-D.'' Two cutoffs were considered for an ''at risk'' classification: the ≥ 16 rule proposed by Radloff (1977) and the ≥ 22 rule used by Cuijpers et al. (2008) for adolescents.
- The curtailed version. This method was performed twice, once for each cutoff. Items were presented in the order that they are listed in Table 1.

- SC via empirical proportions (hereafter SC-Empirical). As with curtailment, this method was performed separately for each cutoff. Three combinations of γ_0 and γ_1 were used: (a) $\gamma_0 = \gamma_1 = 0.90$, (b) $\gamma_0 = \gamma_1 = 0.95$, and (c) $\gamma_0 = \gamma_1 = 0.99$. The stopping rules corresponding to these values will be referred to as SC-Empirical90, SC-Empirical95, and SC-Empirical99, respectively. Items were presented in the order that they are listed in Table 1.
- SC via logistic regression (hereafter SC-Logistic). As with curtailment, this method was performed separately for each cutoff. The same combinations of γ_0 and γ_1 that were used for SC-Empirical were also used for SC-Logistic, resulting in three stopping rules: SC-Logistic90, SC-Logistic95, and SC-Logistic99. Items were presented in the order that they are listed in Table 1.
- CAT. This procedure had previously been validated and applied to the data set as reported in Smits et al. (2011). Briefly, item selection was performed by maximizing the Fisher information at the MAP estimate of θ. Early stopping was performed at stage k if $SE(\hat{\theta}_k) < \varepsilon$. Results from four ε values were tabulated: ε = 0.3, 0.4, 0.5, and 0.6 (Smits et al., 2011, also provide results for ε = 0.7 and 0.8). The methods corresponding to the four ε values will be referred to as CAT-0.3, CAT-0.4, CAT-0.5, and CAT-0.6, respectively. In all cases, an ''at risk'' classification was made if and only if $\hat{\theta}_K \geq \theta^*$. The inverse transformation of the TCC (Equation 5) was used to find the appropriate θ^* value for a given cutoff. The values $X^* = 15.5$ and 21.5 were used in Equation 5, rather than $X^* = 16$ and 22, as a correction for continuity.

The statistical environment R (R Development Core Team, 2005) was used to conduct all simulations.

For the methods requiring training data (SC-Empirical, SC-Logistic, and CAT), it was necessary that performance be evaluated on a different data set from the one used in training. Failure to separate training data from evaluation data can lead to ''capitalization on chance,'' whereby reported results are misleadingly positive (Smits et al., 2011). Therefore, the initial data set of 1,392 respondents was randomly split into two separate halves. In one phase of the simulation, the first half of respondents was used as the training set, and the second half was used in evaluation. In the next phase, the two halves were switched (i.e., the subset that had previously been used for training was used for evaluation, and vice versa). Results from the two phases were then aggregated for reporting.

All methods were evaluated based on the following outcome measures:

1. Respondent burden. The average test length and an *SD* of test lengths were computed. The percentage of respondents whose questionnaires stopped early (i.e., before the 20th item) was also recorded.
2. Concordance with the full-length CES-D. Using the classification from the full-length instrument as a gold standard, the sensitivity and specificity were found.
3. Concordance with the M.I.N.I. Sensitivity and specificity were each calculated twice. In the first analysis, a ''major depression'' diagnosis from the M.I.N.I. was used as the gold standard. In the second analysis, a diagnosis of ''any depression'' (major or minor) from the M.I.N.I. was used as the gold standard.

Lookup tables were also created to express the stopping rules of each curtailed and stochastically curtailed version. Because these tables are provided for comparative purposes and practical usage, rather than the evaluation of methods, there was no need to separate the respondents into

''training'' and ''evaluation'' data sets at this stage. Therefore, to take advantage of all available data, each table was made using the entire set of 1,392 respondents as a ''training'' data set.

Results

Among the 1,392 respondents who took the full-length CES-D, the mean (*SD*) total number-correct score was 13.7 (11.2). In all, 402 respondents (28.9%) were classified as ''at risk'' based on a $\geq$16 cutoff, and 285 respondents (20.5%) were classified as ''at risk'' based on a $\geq$22 cutoff. The coefficient alpha of the full-length scale had previously been reported as .93 (Smits et al., 2011). Among the 242 respondents who received the M.I.N.I., 21 respondents (8.7%) were given a diagnosis of major depression, and 30 respondents (12.4%) were given a diagnosis of any depression (major or minor).

Results When Using a Cutoff of $\geq$16

Table 2 presents the results of each outcome measure for a cutoff of $\geq$16. The full-length CES-D exhibited a sensitivity of 100% and a specificity of 43.4% for predicting a diagnosis of ''major depression'' by the M.I.N.I. The full-length CES-D exhibited a sensitivity of 96.7% and a specificity of 44.8% for predicting a diagnosis of ''any depression'' by the M.I.N.I.

Examining the properties of the sequential stopping rules, curtailment and SC-Empirical99 resulted in an average test length of 16.1 while making identical classifications as the full-length CES-D. SC-Empirical95 and SC-Empirical90 also made the same classification decision as the full-length CES-D for every respondent; the average test lengths of these two methods were 15.8 and 15.4, respectively. SC-Logistic made greater reductions in respondent burden (average test lengths of 9.9, 6.1, and 3.2 for SC-Logistic99, SC-Logistic95, and SC-Logistic90, respectively), but also differed with the full-length CES-D in some classification decisions. Specifically, when using the full-length CES-D as a gold standard, sensitivities (specificities) were 98.5% (99.9%) for SC-Logistic99, 92.0% (99.3%) for SC-Logistic95, and 73.9% (98.9%) for SC-Logistic90. Using the M.I.N.I. diagnoses of ''major'' and ''any'' depression as a gold standard, all SC-Empirical variations had the same sensitivities and specificities as the full-length CES-D (100% sensitivity and 43.4% specificity for major depression, 96.7% sensitivity and 44.8% specificity for any depression). Sensitivities of SC-Logistic ranged from 85.7% to 100% in predicting major depression and from 83.3% to 96.7% in predicting any depression. Specificities of SC-Logistic ranged from 43.9% to 56.6% in predicting major depression and from 45.3% to 58.0% in predicting any depression.

Turning to the CAT results, the average test lengths were 12.7, 7.0, 4.0, and 2.5 for CAT-0.3, CAT-0.4, CAT-0.5, and CAT-0.6, respectively. When using the full-length CES-D as a gold standard, sensitivities (specificities) were 84.1% (96.0%) for CAT-0.3, 75.9% (94.3%) for CAT-0.4, 79.4% (90.4%) for CAT-0.5, and 66.4% (85.9%) for CAT-0.6. Using the M.I.N.I. diagnosis as a gold standard, sensitivities of CAT ranged from 81.0% to 90.5% in predicting major depression and from 76.7% to 86.7% in predicting any depression. Specificities of CAT ranged from 50.7% to 57.5% in predicting major depression and from 51.4% to 58.5% in predicting any depression.

Figure 1 presents side-by-side boxplots of test length by method for the $\geq$16 cutoff. The boxplots demonstrate that the distributions of test length were often not symmetric around the test length means. Several procedures with relatively higher average test lengths (curtailed, SC-Empirical95, and SC-Empirical90) exhibited left-skewed distributions and several procedures with lower average test lengths (SC-Logistic95, SC-Logistic90, CAT-04, CAT-05, and

Table 2. Respondent Burden and Classification Accuracy Results ($\geq$ 16 Cutoff)

		Full-length CES-D	Curtailed[a]	SC-Empirical95	SC-Empirical90	SC-Logistic99	SC-Logistic95	SC-Logistic90	CAT-0.3	CAT-0.4	CAT-0.5	CAT-0.6
Average test length		20.0	16.1	15.8	15.4	9.9	6.1	3.2	12.7	7.0	4.0	2.5
SD of test length		0.0	3.4	3.4	3.4	5.0	4.5	3.6	5.8	5.5	3.9	1.9
% test lengths <20		0.0	89.9	89.9	89.9	93.0	97.6	99.6	68.5	89.4	96.5	100.0
Full-length CES-D as	Sensitivity (%)	100.0	100.0	100.0	100.0	98.5	92.0	73.9	84.1	75.9	79.4	66.4
gold standard	Specificity (%)	100.0	100.0	100.0	100.0	99.9	99.3	98.9	96.0	94.3	90.4	85.9
M.I.N.I. "major depression"	Sensitivity (%)	100.0	100.0	100.0	100.0	100.0	90.5	85.7	90.5	85.7	90.5	81.0
as gold standard	Specificity (%)	43.4	43.4	43.4	43.4	43.9	45.2	56.6	52.0	57.5	50.7	55.2
M.I.N.I. "any depression"	Sensitivity (%)	96.7	96.7	96.7	96.7	96.7	86.7	83.3	86.7	80.0	83.3	76.7
as gold standard	Specificity (%)	44.8	44.8	44.8	44.8	45.3	46.2	58.0	53.3	58.5	51.4	56.1

Note: CES-D = Center for Epidemiologic Studies–Depression; SC = stochastic curtailment; CAT = computerized adaptive testing; M.I.N.I. = Mini-International Neuropsychiatric Interview.
[a]The results for SC-Empirical99 were identical to those of curtailment.

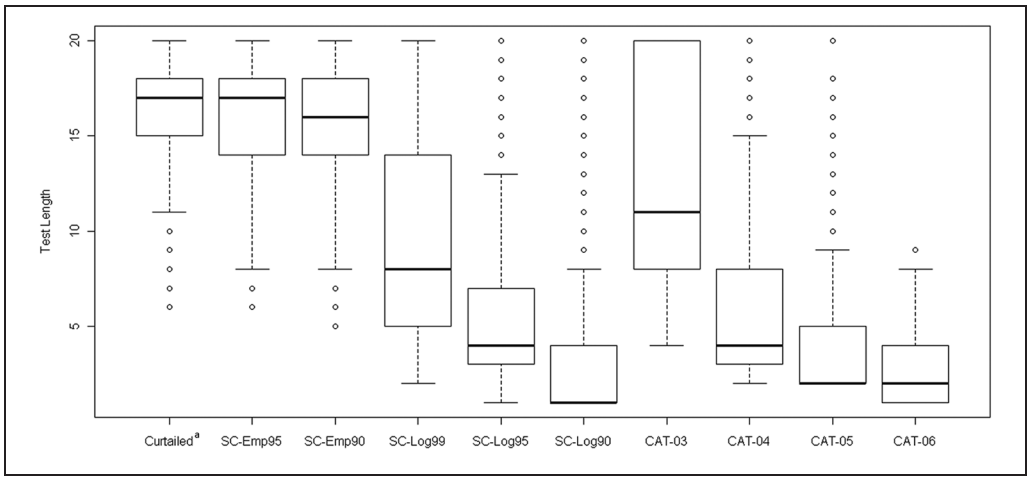

Figure 1. Side-by-side boxplots of test length by method ($\geq$16 cutoff).
Note: SC = stochastic curtailment; CAT = computerized adaptive testing.
[a]The results for SC-Empirical99 were identical to those of curtailment.

CAT-06) exhibited right-skewed distributions. In fact, SC-Logistic90 and CAT-05 had median values that were equal to their respective minimum values.

Results When Using a Cutoff of $\geq$*22*

Table 3 presents the results of each outcome measure for a cutoff of $\geq$22. The full-length CES-D exhibited a sensitivity of 90.5% and a specificity of 66.5% for predicting a diagnosis of ''major depression'' by the M.I.N.I. The full-length CES-D exhibited a sensitivity of 80.0% and a specificity of 67.5% for predicting a diagnosis of ''any depression'' by the M.I.N.I.

Curtailment and SC-Empirical99 had an average test length of 15.6 while making identical classifications as the full-length CES-D. SC-Empirical95 and SC-Empirical90 exhibited average test lengths of 14.8 and 14.3, respectively, and always made the same classification decision as the full-length CES-D. SC-Logistic again made greater reductions in respondent burden than SC-Empirical (average test lengths were 8.3, 4.5, and 3.0 for SC-Logistic99, SC-Logistic95, and SC-Logistic90, respectively) but sometimes differed with the full-length CES-D in terms of classification decisions. Using the full-length CES-D as a gold standard, sensitivities (specificities) were 97.5% (99.9%) for SC-Logistic99, 87.7% (99.5%) for SC-Logistic95, and 78.9% (98.3%) for SC-Logistic90. Using the M.I.N.I. diagnoses of ''major'' and ''any'' depression as a gold standard, all SC-Empirical variations had the same sensitivities and specificities as the full-length CES-D (90.5% sensitivity and 66.5% specificity for major depression, and 80.0% sensitivity and 67.5% specificity for any depression). Sensitivities of SC-Logistic ranged from 85.7% to 90.5% in predicting major depression and from 76.7% to 80.0% in predicting any depression. Specificities of SC-Logistic ranged from 66.5% to 68.8% in predicting major depression and from 67.5% to 69.8% in predicting any depression.

For CAT, the stopping rule was independent of the cutoff; therefore, average test lengths were identical to those presented in the section ''Results When Using a Cutoff of $\geq$16'' (12.7, 7.0, 4.0, and 2.5 for CAT-0.3, CAT-0.4, CAT-0.5, and CAT-0.6, respectively). Using the full-length CES-D as a gold standard, sensitivities (specificities) were 78.9% (98.8%) for CAT-0.3, 76.5% (97.2%) for CAT-0.4, 72.3% (97.7%) for CAT-0.5, and 59.3% (97.5%) for CAT-0.6. Using the M.I.N.I. diagnosis as a gold standard, sensitivities of CAT ranged from 71.4% to

Table 3. Respondent Burden and Classification Accuracy Results ($\geq$22 Cutoff)

		Full-length CES-D	Curtailed[a]	SC-Empirical95	SC-Empirical90	SC-Logistic99	SC-Logistic95	SC-Logistic90	CAT-0.3	CAT-0.4	CAT-0.5	CAT-0.6
Average test length		20.0	15.6	14.8	14.3	8.3	4.5	3.0	12.7	7.0	4.0	2.5
SD of test length		0.0	2.2	2.5	2.6	4.6	4.4	3.3	5.8	5.5	3.9	1.9
% test lengths $<$ 20		0.0	95.8	95.8	95.8	96.8	98.8	99.6	68.5	89.4	96.5	100.0
Full-length CES-D as	Sensitivity (%)	100.0	100.0	100.0	100.0	97.5	87.7	78.9	78.9	76.5	72.3	59.3
gold standard	Specificity (%)	100.0	100.0	100.0	100.0	99.9	99.5	98.3	98.8	97.2	97.7	97.5
M.I.N.I. "major depression" as	Sensitivity (%)	90.5	90.5	90.5	90.5	90.5	90.5	85.7	81.0	71.4	76.2	71.4
gold standard	Specificity (%)	66.5	66.5	66.5	66.5	66.5	68.8	68.8	71.9	71.5	72.9	76.5
M.I.N.I. "any depression" as	Sensitivity (%)	80.0	80.0	80.0	80.0	80.0	80.0	76.7	73.3	63.3	66.7	63.3
gold standard	Specificity (%)	67.5	67.5	67.5	67.5	67.5	69.8	69.8	73.1	72.2	73.6	77.4

Note: CES-D = Center for Epidemiologic Studies–Depression; SC = stochastic curtailment; CAT = computerized adaptive testing; M.I.N.I. = Mini-International Neuropsychiatric Interview.
[a]The results for SC-Empirical99 were identical to those of curtailment.

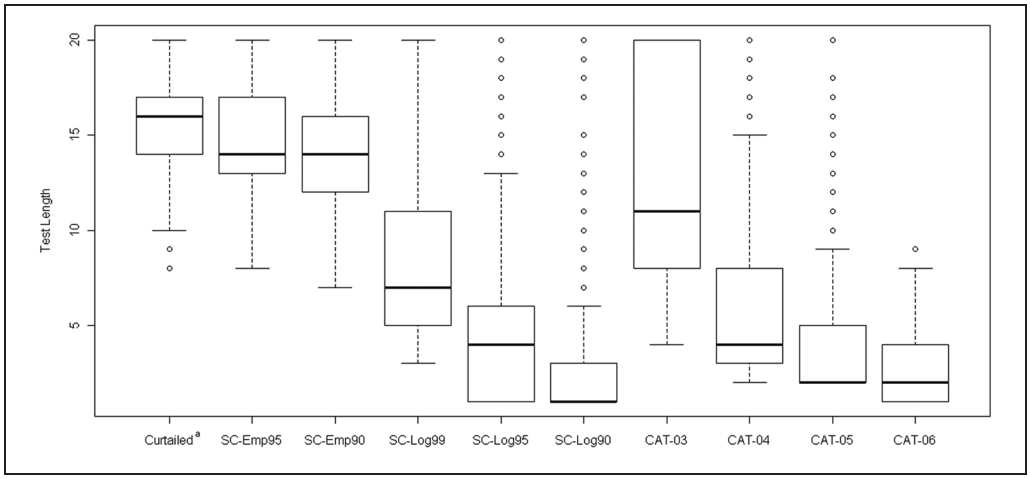

Figure 2. Side-by-side boxplots of test length by method ($\geq$22 cutoff).
Note: SC = stochastic curtailment; CAT = computerized adaptive testing.
[a]The results for SC-Empirical99 were identical to those of curtailment.

81.0% in predicting major depression and from 63.3% to 73.3% in predicting any depression. Specificities of CAT ranged from 71.5% to 76.5% in predicting major depression and from 72.2% to 77.4% in predicting any depression.

Figure 2 displays side-by-side boxplots of test length by method for the $\geq$22 cutoff. As in Figure 1, there were multiple procedures whose test length distributions were not symmetric around their means. SC-Logistic95 and SC-Logistic90 again exhibited prominent right-skewed patterns; the former method's lower quartile was equal to its minimum value, and the latter method's median was equal to its minimum value. The curtailed, SC-Empirical95, and SC-Empirical90 methods exhibited less of a left-skewed pattern for the $\geq$22 cutoff than for the $\geq$16 cutoff, but the first two of these procedures still demonstrated nontrivial asymmetry. The boxplot of each CAT procedure was the same in Figure 2 as in Figure 1.

Stopping Boundaries of Curtailment and SC

Tables 4 and 5 present stopping boundaries corresponding to the rules of curtailment and SC (Table 4 gives results for a $\geq$16 cutoff, whereas Table 5 gives results for a $\geq$22 cutoff). As explained in the section ''A Logistic Regression Approach to SC,'' the tables provide the smallest value of X_k sufficient to stop at time k in favor of an ''at risk'' classification, as well as the largest value of X_k sufficient to stop at time k in favor of a ''not at risk'' classification. Analogous ''lookup'' tables cannot be made for CAT, as this method does not directly use X_k to decide whether to halt testing at time k.

Note that certain stopping boundaries within each table are *vacuous* (i.e., they never arise in light of other boundaries in the table). For example, suppose that a $\geq$16 cutoff has been specified, and SC-Logistic99 is being used as the stopping rule. Table 4 shows that according to the logistic regression analysis, early stopping in favor of an ''at risk'' classification should occur after three items if a new respondent's X_3 value is 6 or higher. The same analysis also reveals that early stopping in favor of an ''at risk'' classification should occur after four items if a new respondent's X_4 value is 9 or higher. Because item scores on the CES-D never exceed 3, however, a respondent's score can only be at least 9 by the fourth stage if it was at least 6 by the third stage. Hence, any respondent whose test would be terminated after Item 4 would have

Table 4. Stopping Boundaries for Curtailment and SC (≥ 16 Cutoff)

	Curtailed[a]		SC-Empirical95		SC-Empirical90		SC-Logistic99		SC-Logistic95		SC-Logistic90	
Items completed (k)	Negative stopping	Positive stopping	Negative stopping	Positive stopping	Negative stopping	Positive stopping	Negative stopping	Positive stopping	Negative stopping	Positive stopping	Negative stopping	Positive stopping
1	NA	NA	NA	NA	NA	NA	NA	NA	NA	$X_k = 3$	$X_k = 0$	$X_k = 3$
2	NA	NA	NA	NA	NA	NA	NA	$X_k = 6$	NA	$X_k \geq 5$	$X_k = 0$[b]	$X_k \geq 4$
3	NA	NA	NA	NA	NA	NA	NA	$X_k \geq 6$	$X_k = 0$	$X_k \geq 5$	$X_k = 0$[b]	$X_k \geq 5$
4	NA	NA	NA	NA	NA	NA	NA	$X_k \geq 9$[b]	$X_k \leq 1$	$X_k \geq 7$	$X_k \leq 1$	$X_k \geq 7$
5	NA	NA	NA	NA	NA	$X_k = 15$	$X_k \leq 1$	$X_k \geq 10$	$X_k \leq 2$	$X_k \geq 8$	$X_k \leq 3$	$X_k \geq 8$
6	NA	$X_k \geq 16$	NA	$X_k \geq 16$	NA	$X_k \geq 15$	$X_k \leq 1$[b]	$Xk \geq 11$	$X_k \leq 3$	$X_k \geq 9$	$X_k \leq 3$[b]	$X_k \geq 8$
7	NA	$X_k \geq 16$	NA	$X_k \geq 16$	NA	$X_k \geq 15$	$X_k \leq 2$	$X_k \geq 11$	$X_k \leq 4$	$X_k \geq 10$	$X_k \leq 4$	$X_k \geq 9$
8	NA	$X_k \geq 16$	NA	$X_k \geq 16$	NA	$X_k \geq 16$	$X_k \leq 3$	$X_k \geq 12$	$X_k \leq 5$	$X_k \geq 11$	$X_k \leq 6$	$X_k \geq 10$
9	NA	$X_k \geq 16$	NA	$X_k \geq 16$	NA	$X_k \geq 16$	$X_k \leq 4$	$X_k \geq 13$	$X_k \leq 5$[b]	$X_k \geq 11$	$X_k \leq 6$[b]	$X_k \geq 11$
10	NA	$X_k \geq 16$	NA	$X_k \geq 16$	NA	$X_k \geq 16$	$X_k \leq 4$[b]	$X_k \geq 13$	$X_k \leq 6$	$X_k \geq 12$	$X_k \leq 6$[b]	$X_k \geq 11$
11	NA	$X_k \geq 16$	NA	$X_k \geq 16$	NA	$X_k \geq 16$	$X_k \leq 5$	$X_k \geq 14$	$X_k \leq 7$	$X_k \geq 13$	$X_k \leq 7$	$X_k \geq 12$
12	NA	$X_k \geq 16$	NA	$X_k \geq 16$	NA	$X_k \geq 16$	$X_k \leq 6$	$X_k \geq 15$	$X_k \leq 8$	$X_k \geq 14$	$X_k \leq 8$	$X_k \geq 13$
13	NA	$X_k \geq 16$	NA	$X_k \geq 16$	$X_k = 0$	$X_k \geq 16$	$X_k \leq 7$	$X_k \geq 16$	$X_k \leq 9$	$X_k \geq 14$	$X_k \leq 9$	$X_k \geq 14$
14	NA	$X_k \geq 16$	$X_k \leq 1$	$X_k \geq 16$	$X_k \leq 2$	$X_k \geq 16$	$X_k \leq 8$	$X_k \geq 16$	$X_k \leq 9$[b]	$X_k \geq 15$	$X_k \leq 10$	$X_k \geq 14$
15	$X_k = 0$	$X_k \geq 16$	$X_k \leq 2$	$X_k \geq 16$	$X_k \leq 3$	$X_k \geq 16$	$X_k \leq 9$	$X_k \geq 16$	$X_k \leq 10$	$X_k \geq 15$	$X_k \leq 11$	$X_k \geq 15$
16	$X_k \leq 3$	$X_k \geq 16$	$X_k \leq 5$	$X_k \geq 16$	$X_k \leq 6$	$X_k \geq 16$	$X_k \leq 10$	$X_k \geq 16$	$X_k \leq 11$	$X_k \geq 16$	$X_k \leq 12$	$X_k \geq 16$
17	$X_k \leq 6$	$X_k \geq 16$	$X_k \leq 7$	$X_k \geq 16$	$X_k \leq 8$	$X_k \geq 16$	$X_k \leq 10$[b]	$X_k \geq 16$	$X_k \leq 12$	$X_k \geq 16$	$X_k \leq 12$[b]	$X_k \geq 16$
18	$X_k \leq 9$	$X_k \geq 16$	$X_k \leq 10$	$X_k \geq 16$	$X_k \leq 10$	$X_k \geq 16$	$X_k \leq 11$	$X_k \geq 16$	$X_k \leq 12$[b]	$X_k \geq 16$	$X_k \leq 13$	$X_k \geq 16$
19	$X_k \leq 12$	$X_k \geq 16$	$X_k \leq 12$	$X_k \geq 16$	$X_k \leq 12$	$X_k \geq 16$	$X_k \leq 12$	$X_k \geq 16$	$X_k \leq 13$	$X_k \geq 16$	$X_k \leq 13$[b]	$X_k \geq 16$
20	$X_k \leq 15$	$X_k \geq 16$	$X_k \leq 15$	$X_k \geq 16$	$X_k \leq 15$	$X_k \geq 16$	$X_k \leq 15$	$X_k \geq 16$	$X_k \leq 15$	$X_k \geq 16$	$X_k \leq 15$	$X_k \geq 16$

Note: SC = stochastic curtailment.

[a]The boundaries for the curtailed version are also the boundaries for SC-Empirical99.

[b]Stopping boundary is vacuous.

Table 5. Stopping Boundaries for Curtailment and SC (≥ 22 Cutoff)

	Curtailed[a]		SC-Empirical95		SC-Empirical90		SC-Logistic99		SC-Logistic95		SC-Logistic90	
Items completed (k)	Negative stopping	Positive stopping	Negative stopping	Positive stopping	Negative stopping	Positive stopping	Negative stopping	Positive stopping	Negative stopping	Positive stopping	Negative stopping	Positive stopping
1	NA	NA	NA	NA	NA	NA	NA	NA	$X_k = 0$	NA	$X_k = 0$	$X_k = 3$
2	NA	NA	NA	NA	NA	NA	NA	NA	$X_k = 0$[b]	$X_k = 6$	$X_k = 0$[b]	$X_k \geq 5$
3	NA	NA	NA	NA	NA	NA	NA	$X_k \geq 7$	$X_k = 0$[b]	$X_k \geq 6$	$X_k \leq 1$	$X_k \geq 6$
4	NA	NA	NA	NA	NA	NA	$X_k = 0$	$X_k \geq 10$[b]	$X_k \leq 2$	$X_k \geq 8$	$X_k \leq 2$	$X_k \geq 7$
5	NA	NA	NA	NA	NA	NA	$X_k \leq 2$	$X_k \geq 11$	$X_k \leq 3$	$X_k \geq 10$	$X_k \leq 4$	$X_k \geq 9$
6	NA	NA	NA	NA	NA	NA	$X_k \leq 2$[b]	$Xk \geq 12$	$X_k \leq 4$	$X_k \geq 11$	$X_k \leq 5$	$X_k \geq 10$
7	NA	NA	NA	NA	NA	$X_k = 21$	$X_k \leq 3$	$X_k \geq 13$	$X_k \leq 5$	$X_k \geq 12$	$X_k \leq 6$	$X_k \geq 11$
8	NA	$X_k \geq 22$	NA	$X_k \geq 22$	NA	$X_k \geq 22$	$X_k \leq 5$	$X_k \geq 15$	$X_k \leq 6$	$X_k \geq 13$	$X_k \leq 7$	$X_k \geq 12$
9	NA	$X_k \geq 22$	NA	$X_k \geq 22$	NA	$X_k \geq 22$	$X_k \leq 5$[b]	$X_k \geq 15$	$X_k \leq 7$	$X_k \geq 14$	$X_k \leq 8$	$X_k \geq 13$
10	NA	$X_k \geq 22$	NA	$X_k \geq 22$	NA	$X_k \geq 22$	$X_k \leq 6$	$X_k \geq 16$	$X_k \leq 8$	$X_k \geq 14$	$X_k \leq 8$[b]	$X_k \geq 14$
11	NA	$X_k \geq 22$	NA	$X_k \geq 22$	$X_k = 0$	$X_k \geq 22$	$X_k \leq 7$	$X_k \geq 17$	$X_k \leq 9$	$X_k \geq 16$	$X_k \leq 10$	$X_k \geq 15$
12	NA	$X_k \geq 22$	$X_k = 0$	$X_k \geq 22$	$X_k \leq 3$	$X_k \geq 22$	$X_k \leq 9$	$X_k \geq 18$	$X_k \leq 10$	$X_k \geq 17$	$X_k \leq 11$	$X_k \geq 16$
13	$X_k = 0$	$X_k \geq 22$	$X_k \leq 3$	$X_k \geq 22$	$X_k \leq 4$	$X_k \geq 22$	$X_k \leq 10$	$X_k \geq 19$	$X_k \leq 11$	$X_k \geq 18$	$X_k \leq 12$	$X_k \geq 17$
14	$X_k \leq 3$	$X_k \geq 22$	$X_k \leq 6$	$X_k \geq 22$	$X_k \leq 7$	$X_k \geq 22$	$X_k \leq 11$	$X_k \geq 20$	$X_k \leq 13$	$X_k \geq 18$	$X_k \leq 13$	$X_k \geq 18$
15	$X_k \leq 6$	$X_k \geq 22$	$X_k \leq 7$	$X_k \geq 22$	$X_k \leq 9$	$X_k \geq 22$	$X_k \leq 13$	$X_k \geq 20$	$X_k \leq 14$	$X_k \geq 19$	$X_k \leq 15$	$X_k \geq 18$
16	$X_k \leq 9$	$X_k \geq 22$	$X_k \leq 10$	$X_k \geq 22$	$X_k \leq 11$	$X_k \geq 22$	$X_k \leq 14$	$X_k \geq 21$	$X_k \leq 15$	$X_k \geq 20$	$X_k \leq 16$	$X_k \geq 19$
17	$X_k \leq 12$	$X_k \geq 22$	$X_k \leq 13$	$X_k \geq 22$	$X_k \leq 14$	$X_k \geq 22$	$X_k \leq 15$	$X_k \geq 22$	$X_k \leq 16$	$X_k \geq 21$	$X_k \leq 16$[b]	$X_k \geq 20$
18	$X_k \leq 15$	$X_k \geq 22$	$X_k \leq 15$	$X_k \geq 22$	$X_k \leq 16$	$X_k \geq 22$	$X_k \leq 17$	$X_k \geq 22$	$X_k \leq 18$	$X_k \geq 21$	$X_k \leq 18$	$X_k \geq 21$
19	$X_k \leq 18$	$X_k \geq 22$	$X_k \leq 18$	$X_k \geq 22$	$X_k \leq 18$	$X_k \geq 22$	$X_k \leq 18$	$X_k \geq 22$	$X_k \leq 18$[b]	$X_k \geq 22$	$X_k \leq 19$	$X_k \geq 22$
20	$X_k \leq 21$	$X_k \geq 22$	$X_k \leq 21$	$X_k \geq 22$	$X_k \leq 21$	$X_k \geq 22$	$X_k \leq 21$	$X_k \geq 22$	$X_k \leq 21$	$X_k \geq 22$	$X_k \leq 21$	$X_k \geq 22$

Note: SC = stochastic curtailment.

[a]The boundaries for the curtailed version are also the boundaries for SC-Empirical99.

[b]Stopping boundary is vacuous.

already had it terminated after Item 3, so the boundary for Item 4 can never be invoked. Similarly, because CES-D item scores are never negative, and an $X_5 \leq 1$ boundary exists for the fifth stage, SC-Logistic99's boundary of $X_6 \leq 1$ for the sixth stage can never be invoked. Such vacuous boundaries have no impact on the statistical properties of SC, but their existence is mentioned for purposes of completeness.

Summary and Discussion

Although the CES-D has been studied for decades in a wide variety of settings (Sephton et al., 2009), it has only recently been paired with modern methodology for computer-based administration. In particular, Smits et al. (2011) showed that CAT can reduce the respondent burden associated with taking the CES-D. However, the sequential analysis techniques of curtailment and SC had never been studied alongside this instrument. Therefore, the purposes of the present research were to (a) adapt previous formulations of curtailment and SC to the CES-D; (b) introduce the ''logistic regression formulation'' of SC, which can be used with other health questionnaires in addition to the CES-D; and (c) compare all procedures (curtailment, SC, CAT, and the full-length CES-D) in terms of their classification properties and average test lengths, using a post hoc simulation of real data.

Results indicated that the sequential and adaptive procedures have potential to economize the administration of CES-D items in a classification setting. A simple curtailment rule reduced the respondent burden by as much as 22% while always making the same classification as the full-length CES-D. SC-Empirical (specifically, SC-Empirical90) increased this percentage to 23% for the $\geq$16 cutoff and 28% for the $\geq$22 cutoff. SC-Logistic and CAT made substantially greater gains in respondent burden while sometimes diverging from the classification decision of the full-length test.

Making comparisons between methods is challenging because two evaluation criteria were considered—test length and concordance with a ''gold standard'' classification—and the method that performs best with respect to one criterion might not perform the best with respect to the other criterion. Veldkamp (1999) discussed the similar problem of optimizing multiple objectives in a general test assembly context; several authors have considered multiple objectives within the classification context (Lewis & Sheehan, 1990; Spray & Reckase, 1996; Vos, 2000). One approach is to match the different procedures (here, CAT and curtailment/SC) based on concordance and compare them with respect to test length, or vice versa (Thompson, 2011). In the present context, it is difficult to match via concordance due to the fact that there are multiple numbers related to this criterion (namely, the sensitivity and specificity with respect to each gold standard). Therefore, the procedures were matched based on a single number–average test length–and compared with respect to concordance. For the $\geq$16 cutoff, SC-Logistic95 and CAT-0.4 were close with regard to average test length (6.1 vs. 7.0, respectively), as were SC-Logistic90 and CAT-0.6 (3.2 vs. 2.5, respectively). Comparing these methods' classification properties, the SC-Logistic procedures exhibited greater concordance with the full-length CES-D (higher sensitivity and specificity than the CAT method with similar average test length). The SC-Logistic procedures also exhibited higher sensitivity in predicting the M.I.N.I. diagnosis. CAT-0.4 had higher specificity than SC-Logistic95 in predicting the M.I.N.I. diagnosis; SC-Logistic90 had slightly higher specificity than CAT-0.6. For the $\geq$22 cutoff, SC-Logistic90 was close to CAT-0.6 with regard to average test length (3.0 vs. 2.5, respectively), and SC-Logistic95 was close to CAT-0.5 (4.5 vs. 4.0, respectively). The SC-Logistic procedures again exhibited greater concordance with the full-length CES-D (higher sensitivity and specificity than the CAT method with similar average test length). The SC-Logistic procedures

also exhibited higher sensitivity in predicting the M.I.N.I. diagnoses, whereas the CAT procedures exhibited higher specificity for them.

Which method to use for a given application depends on the goal of the practitioner, as well as the relative importance of respondent burden versus concordance with the full-length test. When a practitioner is using the CES-D for *estimation* rather than *classification* of respondents, CAT is more appropriate than curtailment and SC (which are both specifically tailored to classification). Another CAT advantage is that it allows the practitioner to understand the properties of the test items and make comparisons between them. An additional potential benefit of CAT is that if the IRT model fits the data well, the IRT-CAT combination could theoretically result in more accurate classifications than the scoring method of the section 'Traditional Scoring of the CES-D.' However, Smits et al. (2011) did not find an empirical basis for this phenomenon. Turning to curtailment, benefits of this method are as follows: (a) It is guaranteed to make the same classification decision as the full-length CES-D, while sometimes reducing the respondent burden; (b) it does not require the existence of training data to be applied to new respondents; and (c) it is simple to understand and use (the stopping boundaries are identical for every respondent and are easier to develop and program than IRT–CAT). Finally, benefits of SC are as follows: (a) It results in lower average test lengths than curtailment, (b) it displays more concordance with the full-length CES-D than a CAT of comparable respondent burden, and (c) once the appropriate calculations from the training data have been performed, SC is as easy to apply as curtailment (its stopping boundaries can be written as a lookup table).

If the decision to use SC is made, one important consideration is the values of γ_0 and γ_1 to use within this procedure. M. D. Finkelman et al. (2011) used $\gamma_0 = \gamma_1 = 0.95$ in their simulations, but the value of 0.95 was chosen arbitrarily. For the present article, an approach is adopted similar to how Thompson (2011) selected between different versions of a variable-length test. Specifically, it is recommended that γ_0 and γ_1 be selected to achieve the greatest reduction in average test length, among the set of γ_0 and γ_1 values for which concordance with the full-length CES-D is within an acceptable tolerance. Under this framework, SC-Empirical90 was preferable to SC-Empirical95 and SC-Empirical99 in the present study: All three of these procedures exhibited perfect concordance with the full-length CES-D, but SC-Empirical90 had the lowest average test length among them. However, a practitioner might select SC-Logistic99 over SC-Logistic95 and SC-Logistic90: The former was the only method that always exhibited more than 90% sensitivity and 90% specificity for the full-length CES-D while still reducing the average test length by more than 50%. Future applications of SC should likewise include experimentation with different γ_0 and γ_1 values to find the appropriate balance between test length and classification concordance.

The decision of whether to use SC-Empirical or SC-Logistic in a given application may hinge not only on average test lengths and classification properties but also on logistic regression's level of fit in the data. Unlike the nonparametric estimation process used by SC-Empirical, estimation of $\hat{P}_k^l$ by SC-Logistic involves parametric modeling and thus has greater reliance on adequate goodness of fit. The Hosmer–Lemeshow test (Hosmer & Lemeshow, 1989) may be used to assess the fit of a logistic regression model, and p-values less than a prespecified threshold indicate statistically significant misfit. In cases where there is significant evidence of misfit, logistic regression may lead to misleading conclusions and SC-Empirical may be preferred instead.

Comparing the results of this study to those of other researchers, the reductions in average test length achieved by CAT herein were in the range of previous work (Gibbons et al., 2008; Moreno, Wetzel, McBride, & Weiss, 1984; Weiss, 1982). The reductions derived from curtailment and SC-Empirical were more modest than those found when M. D. Finkelman et al. (2011) applied these procedures to the Medicare Health Outcomes Survey. The latter finding is

likely due to the fact that in the scoring method of the CES-D described in the section Traditional Scoring of the CES-D, all items receive equal weight in the total score, whereas in the scoring of the Medicare Health Outcomes Survey, M. D. Finkelman et al. gave different weights to different items on the questionnaire. By placing the least influential items at the end of the test, and thereby avoiding the presentation of such items via early stopping, M. D. Finkelman et al. increased the gains of curtailment and SC. However, the use of SC-Logistic in the present study resulted in even larger savings of items than M. D. Finkelman et al. found with SC-Empirical. Because SC-Logistic was not used in M. D. Finkelman et al., however, and had never been proposed prior to this study, no ''apples to apples'' comparison of its results herein can be made with previous work.

The purpose of providing specific stopping boundaries in Tables 4 and 5 was to allow investigators to use them in future administrations of the CES-D. The curtailment boundaries can be applied to *all* future CES-D respondents, as long as the cutoff is $\geq$16 (Table 4) or $\geq$22 (Table 5). Analogous boundaries can easily be calculated for other cutoff values by following the general rule provided in the section ''Curtailment.'' By contrast, caution must be exercised when considering the use of the SC boundaries presented in Tables 4 and 5. In particular, the SC boundaries are dependent on the training sample; hence, they should not be applied to respondents for whom the sample of this study was not representative. If SC (either SC-Empirical or SC-Logistic) is to be used for a new population, stopping boundaries should be created by training the method to data from that same population.

This article represents a first step toward the use of curtailment and SC in live CES-D administration. However, several important limitations are notable. First, as alluded to earlier, the results of this study are based on one adolescent sample that is unlikely to be representative of other populations taking the CES-D. The classification properties and the average test length of a given method may differ when applied to other respondent groups. Second, the adaptive and sequential rules were applied post hoc rather than being used in an operational setting. It is possible that the results could change between simulated and actual administrations, although previous research (e.g., Kocalevent et al., 2009) suggests that such an effect is likely to be slight. As explained by Smits et al. (2011), the programming, implementation, and maintenance of adaptive and sequential techniques are also more complex in actual administration than in simulation, especially when respondents are assessed over the Internet. Therefore, further research is needed before these methods can be operationalized. Additional post hoc simulations should be performed, using diverse populations of respondents, to compare CAT, curtailment, and SC to one another—as well as to other CCT methods, such as the SPRT, IRT-based confidence intervals, and decision theory. Results under different conditions (e.g., different values of X^*, γ_0, and γ_1) should be examined. Finally, all methods should be pilot tested to obtain feedback from respondents, to test the practicality of their implementation by computer, and to evaluate their efficiency in live settings. Each of these undertakings will be addressed in future work.

Acknowledgments

The authors would like to thank Pim Cuijpers for providing the data. They are also indebted to Tofool Al Ghanem, Eric Bourke, Ronald Kulich, and two anonymous reviewers for their helpful comments.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Adams, L. M., & Gale, D. (1982). Solving the quandary between questionnaire length and response rate in educational research. *Research in Higher Education, 17*, 231-240.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Bartroff, J., Finkelman, M., & Lai, T. L. (2008). Modern sequential analysis and its applications to computerized adaptive testing. *Psychometrika, 73*, 473-486.

Breslau, N. (1985). Depressive symptoms, major depression, and generalized anxiety: A comparison of self-reports on CES-D and results from diagnostic interviews. *Psychiatry Research, 15*, 219-229.

Carpenter, J. S., Andrykowski, M. A., Wilson, J., Hall, L. A., Rayens, M. K., Sachs, B., & Cunningham, L. L. (1998). Psychometrics for two short forms of the Center for Epidemiologic Studies—Depression Scale. *Issues in Mental Health Nursing, 19*, 481-494.

Chabrol, H., Montovany, A., Chouicha, K., & Duconge, E. (2002). Study of the CES-D on a sample of 1,953 adolescent students. *Encephale, 28*, 429-432.

Chang, H.-H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33*, 333-353.

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*, 213-229.

Chang, Y.-C. I. (2005). Application of sequential interval estimation to adaptive mastery testing. *Psychometrika, 70*, 685-713.

Choi, S. W., Reise, S. P., Pilkonis, P. A., Hays, R. D., & Cella, D. (2010). Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. *Quality of Life Research, 19*, 125-136.

Cole, J. C., Rabin, A. S., Smith, T. L., & Kaufman, A. S. (2004). Development and validation of a Rasch-derived CES-D short form. *Psychological Assessment, 16*, 360-372.

Compton, W. M., Conway, K. P., Stinson, F. S., & Grant, B. F. (2006). Changes in the prevalence of major depression and comorbid substance use disorders in the United States between 1991-1992 and 2001-2002. *American Journal of Psychiatry, 163*, 2141-2147.

Crane, P. K., Gibbons, L. E., Jolley, L., & van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques: DIFdetect and difwithpar. *Medical Care, 44*, S115-S123.

Cuijpers, P., Boluijt, P., & van Straten, A. (2008). Screening of depression in adolescents through the Internet: Sensitivity and specificity of two screening questionnaires. *European Child & Adolescent Psychiatry, 17*, 32-38.

Davis, B. R., & Hardy, R. J. (1994). Data monitoring in clinical trials: The case for stochastic curtailment. *Journal of Clinical Epidemiology, 47*, 1033-1042.

Dodd, B. G., De Ayala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 19*, 5-22.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23*, 249-261.

Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement, 60*, 713-734.

Eisenberg, B., & Ghosh, B. K. (1980). Curtailed and uniformly most powerful sequential tests. *Annals of Statistics, 8*, 1123-1131.

Eisenberg, B., & Simons, G. (1978). On weak admissibility of tests. *Annals of Statistics, 6*, 319-332.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Fechner-Bates, S., Coyne, J. C., & Schwenk, T. L. (1994). The relationship of self-reported distress to depressive disorders and other psychopathology. *Journal of Consulting and Clinical Psychology, 62*, 550-559.

Finkelman, M. (2008). On using stochastic curtailment to shorten the SPRT in sequential mastery testing. *Journal of Educational and Behavioral Statistics, 33*, 442-463.

Finkelman, M. D. (2010). Variations on stochastic curtailment in sequential mastery testing. *Applied Psychological Measurement, 34*, 27-45.

Finkelman, M. D., He, Y., Kim, W., & Lai, A. M. (2011). Stochastic curtailment of health questionnaires: A method to reduce respondent burden. *Statistics in Medicine, 30*, 1989-2004.

Fries, J. F., Cella, D., Rose, M., Krishnan, E., & Bruce, B. (2009). Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing. *Journal of Rheumatology, 36*, 2061-2066.

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., . . . Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services, 59*, 361-368.

Glassman, A. H., & Shapiro, P. A. (1998). Depression and the course of coronary artery disease. *American Journal of Psychiatry, 155*, 4-11.

Grzywacz, J. G., Hovey, J. D., Seligman, L. D., Arcury, T. A., & Quandt, S. A. (2006). Evaluating short-form versions of the CES-D for measuring depressive symptoms among immigrants from Mexico. *Hispanic Journal of Behavioral Sciences, 28*, 404-424.

Hambleton, R. K., & de Gruijter, D. N. M. (1983). Application of item response models to criterion-referenced test item selection. *Journal of Educational Measurement, 20*, 355-367.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: SAGE.

Herzog, A. R., & Bachman, J. G. (1980). Effects of questionnaire length on response quality. *Public Opinion Quarterly, 45*, 549-559.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression*. New York, NY: Wiley.

Kessler, R. C., McGonagle, K. A., Zhao, S., Nelson, C. B., Hughes, M., Eshelman, S., . . . Kendler, K. S. (1994). Lifetime and 12-month prevalence of *DSM-III-R* psychiatric disorders in the United States: Results from the National Comorbidity Study. *Archives of General Psychiatry, 51*, 8-19.

Kocalevent, R. D., Rose, M., Becker, J., Walter, O. B., Fliege, H., Bjorner, J. B., . . . Klapp, B. F. (2009). An evaluation of patient-reported outcomes found computerized adaptive testing was efficient in assessing stress perception. *Journal of Clinical Epidemiology, 62*, 278-287.

Kohout, F. J., Berkman, L. F., Evans, D. A., & Cornoni-Huntley, J. (1993). Two shorter forms of the CES-D depression symptoms index. *Journal of Aging and Health, 5*, 179-193.

Lan, K. K. G., Simon, R., & Halperin, M. (1982). Stochastically curtailed tests in long-term clinical trials. *Communications in Statistics-Sequential Analysis, 1*, 207-219.

Leung, D. H.-Y., Wang, Y.-G., & Amar, D. (2003). Early stopping by using stochastic curtailment in a three-arm sequential trial. *Applied Statistics, 52*, 139-152.

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14*, 367-386.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9*, 49-57.

Mitchell, A. J., & Coyne, J. C. (2007). Do ultra-short screening instruments accurately detect depression in primary care? A pooled analysis and meta-analysis of 22 studies. *British Journal of General Practice, 57*, 144-151.

Mojarrad, Y., & Lennings, C. J. (2002). Examination of the Centre for Epidemiological Studies Depression Scale (CES-D) in an adolescent mental health sample. *Journal of Applied Health Behaviour, 4*, 1-6.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague, the Netherlands: Mouton.

Moreno, K. E., Wetzel, C. D., McBride, J. R., & Weiss, D. J. (1984). Relationship between corresponding Armed Services Vocational Aptitude Battery and computerized adaptive testing subtests. *Applied Psychological Measurement, 8*, 155-163.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Murray, C. J. L., & Lopez, A. D. (Eds.). (1996). *The global burden of disease: A comprehensive assessment of mortality and disability from diseases, injuries and risk factors in 1990 and projected to 2020*. Cambridge, MA: Harvard University Press.

Myers, J. K., & Weissman, M. M. (1980). Use of a self-report symptom scale to detect depression in a community sample. *American Journal of Psychiatry, 137*, 1081-1084.

Pandya, R., Metz, L., & Patten, S. B. (2005). Predictive value of the CES-D in detecting depression among candidates for disease-modifying multiple sclerosis treatment. *Psychosomatics, 46*, 131-134.

Poulin, C., Hand, D., & Boudreau, B. (2005). Validity of a 12-item version of the CES-D used in the national longitudinal study of children and youth. *Chronic Diseases in Canada, 26*, 65-72.

Prescott, C. A., McArdle, J. J., Hishinuma, E. S., Johnson, R. C., Miyamoto, R. H., Andrade, N. N., . . . Carlton, B. S. (1998). Prediction of major depression and dysthymia from CES-D scores among ethnic minority adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry, 37*, 495-503.

Radloff, L. S. (1977). The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*, 385-401.

R Development Core Team. (2005). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from http://www.R-project.org

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 237-255). New York, NY: Academic Press.

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., . . . PROMIS Cooperative Group. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care, 45*, S22-S31.

Roberts, R. E., Lewinsohn, P. M., & Seeley, J. R. (1991). Screening for adolescent depression: A comparison of depression scales. *Journal of the American Academy of Child & Adolescent Psychiatry, 30*, 58-66.

Roberts, R. E., & Vernon, S. W. (1983). The Center for Epidemiologic Studies Depression Scale: Its use in a community sample. *American Journal of Psychiatry, 140*, 41-46.

Rudner, L. M. (2009). Scoring and classifying examinees using measurement decision theory. *Practical Assessment, Research & Evaluation, 14*. Retrieved from http://pareonline.net/getvn.asp?v=14&n=8

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores [Monograph Supplement No. 17]. *Psychometrika*.

Santor, D. A., & Coyne, J. C. (1997). Shortening the CES–D to improve its ability to detect cases of depression. *Psychological Assessment, 9*, 233-243.

Scientific Advisory Committee of the Medical Outcomes Trust. (2002). Assessing health status and quality-of-life instruments: Attributes and review criteria. *Quality of Life Research, 11*, 193-205.

Sephton, S. E., Dhabhar, F. S., Keuroghlian, A. S., Giese-Davis, J., McEwen, B. S., Ionan, A. C., & Spiegel, D. (2009). Depression, cortisol, and suppressed cell-mediated immunity in metastatic breast cancer. *Brain, Behavior, and Immunity, 23*, 1148-1155.

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., . . . Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I.): The development and validation of a structured diagnostic psychiatric interview for *DSM-IV* and ICD-10. *Journal of Clinical Psychiatry, 59*(Suppl. 20), 22-57.

Smits, N., Cuijpers, P., & van Straten, A. (2011). Applying computerized adaptive testing to the CES-D scale: A simulation study. *Psychiatry Research, 188*, 147-155.

Smits, N., Zitman, F. G., Cuijpers, P., den Hollander-Gijsman, M. E., & Carlier, I. V. E. (2012). A proof of principle for using adaptive testing in routine outcome monitoring: The efficiency of the Mood and

Anxiety Symptoms Questionnaire–Anhedonic Depression CAT. *BMC Medical Research Methodology, 12*. Retrieved form http://www.biomedcentral.com/1471-2288/12/4

Snapinn, S., Chen, M.-G., Jiang, Q., & Koutsoukos, T. (2006). Assessment of futility in clinical trials. *Pharmaceutical Statistics, 5*, 273-281.

Spitzer, R. L., Kroenke, K., Linzer, M., Hahn, S. R., Williams, J. B. W., DeGruy, F. V., . . . Davies, M. (1995). Health-related quality of life in primary care patients with mental disorders. Results from the PRIME-MD 1000 study. *Journal of the American Medical Association, 274*, 1511-1517.

Spray, J. A., & Reckase, M. D. (1996). Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics, 21*, 405-414.

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 101-133). Mahwah, NJ: Erlbaum.

Thompson, N. A. (2007). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment, Research & Evaluation, 12*. Retrieved from http://pareonline.net/getvn.asp?v=12&n=1

Thompson, N. A. (2011). Termination criteria for computerized classification testing. *Practical Assessment, Research & Evaluation, 16*. Retrieved from http://pareonline.net/getvn.asp?v=16&n=4

U.S. Preventive Services Task Force. (2002). Screening for depression: Recommendations and rationale. *Annals of Internal Medicine, 136*, 760-764.

Veldkamp, B. P. (1999). Multiple objective test assembly problems. *Journal of Educational Measurement, 36*, 253-266.

Vos, H. J. (2000). A Bayesian procedure in the context of sequential mastery testing. *Psicológica, 21*, 191-211.

Wald, A. (1947). *Sequential analysis*. New York, NY: Wiley.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*, 473-492.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361-375.

Weissman, A. (2007). Mutual information item selection in adaptive classification testing. *Educational and Psychological Measurement, 67*, 41-58.

Weissman, M. M., Sholomskas, D., Pottenger, M., Prusoff, B. A., & Locke, B. Z. (1977). Assessing depressive symptoms in five psychiatric populations: A validation study. *American Journal of Epidemiology, 106*, 203-214.

Wells, K. B., Sturm, R., Sherbourne, C. D., & Meredith, L. S. (1996). *Caring for depression*. Cambridge, MA: Harvard University Press.

World Health Organization. (2001). *Mental health: New understanding, new hope*. Geneva, Switzerland: Author.

Yang, H. J., Soong, W. T., Kuo, P. H., Chang, H. L., & Chen, W. J. (2004). Using the CES-D in a two-phase survey for depressive disorders among nonreferred adolescents in Taipei: A stratum-specific likelihood ratio analysis. *Journal of Affective Disorders, 82*, 419-430.